medRxiv preprint doi: https://doi.org/10.64898/2026.04.13.26349911; this version posted April 14, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

Leveraging State-of-the-Art LLMs for the De-identification of Sensitive Health Information in Clinical Speech.

Hong-Jie Dai [1,2,3,4], Tatheer Hussain Mir [1], Liang-Chun Fang [1], Ching-Tai Chen [5], Hui-Hsien Feng [6], Jiunn-Ru Lai [7], Hsieh-Chih Hsu [1], Pratham Nandy [8], Omkar Panchal [8], Wei-Hsiang Liao [9], You-Zhen Tien [9], Pin-Zhen Chen [9], Yun-Ru Lin [9], Jitendra Jonnagaddala [10, 11*]

[1] Intelligent System Laboratory, Department of Electrical Engineering, College of Electrical Engineering and Computer Science, National Kaohsiung University of Science and Technology, Kaohsiung, Taiwan

[2] National Institute of Cancer Research, National Health Research Institutes, Tainan, Taiwan
[3] School of Post-Baccalaureate Medicine, College of Medicine, Kaohsiung Medical University, Kaohsiung, Taiwan
[4] Center for Big Data Research, Kaohsiung Medical University, Kaohsiung, Taiwan
[5] Machine Learning and Bioinformatics Lab, Department of Computer Science, University of Taipei, Taipei, Taiwan

[6] Department of English, National Kaohsiung University of Science and Technology, Kaohsiung, Taiwan
[7] Wireless Networking and Mobile Computing Research Lab, Department of Electrical Engineering, College of Electrical Engineering and Computer Science, National Kaohsiung University of Science and Technology, Kaohsiung, Taiwan
[8] CGD Health Pvt. Ltd, Mumbai, India
[9] Department of Bioinformatics and Medical Engineering, Asia University, Taichung, Taiwan
[10] Discipline of General Practice, School of Clinical Medicine, UNSW Sydney, Kensington, Australia
[11] SREDH Consortium, Sydney, Australia

[*] *Corresponding author*

Dr. Jitendra Jonnagaddala; Discipline of General Practice, School of Clinical Medicine, UNSW Sydney, Kensington NSW 2052; z3339253@unsw.edu.au

Abstract

Accurate recognition and deidentification of sensitive health information (SHI) in spoken dialogues requires multimodal algorithms that can understand medical language and contextual nuance. However, the recognition and deidentification risks expose sensitive health information (SHI). Additionally, the variability and complexity of medical terminology, along with the inherent biases in medical datasets, further complicate this task. This study introduces the SREDH/AI-Cup 2025 Medical Speech Sensitive Information Recognition Challenge, which focuses on two tasks: Task-1: Speech transcription systems must accurately transcribe speech into text; and Task-2: Medical speech de-identification to detect and appropriately classify mentions of SHI. The competition attracted 246 teams; top-performing systems

NOTE: This preprint reports new research that has not been certified by peer review and should not be used to guide clinical practice.

medRxiv preprint doi: https://doi.org/10.64898/2026.04.13.26349911; this version posted April 14, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

achieved a mixed error rate (MER) of 0.1147 and a macro F1-score of 0.7103, with average MER and macro F1-score of 0.3539 and 0.2696, respectively. Results were presented at the IW-DMRN workshop in 2025. Notably, the results reveal that LLMs were prevalent across both tasks: 97.5% of teams adopted LLMs for Task 1 and 100% for Task 2. Highlighting their growing role in healthcare. Furthermore, we finetuned six models, demonstrating strong precision (~0.885–0.889) with slightly lower recall (~0.830–0.847), resulting in F1-scores of 0.857–0.867.

Keywords: Speech Deidentification, Large Language Models, Deidentification, Sensitive Health Information, Natural Language Processing.

Word count: Abstract: 187; Main text: 4012 (excluding abstract, tables, figures and references)

Introduction

AI-driven health systems analyze large volumes of clinical data from diverse sources, including structured and unstructured clinical text such as free-text clinical notes and electronic health records (EHRs). The data contains valuable information rich in semantic and contextual cues, used for downstream analysis and secondary use, including clinical studies (1), clinical audits for quality enhancement (2), identification of social determinants of health (SDoH) (3), public health surveillance (4), and improvement of healthcare practices (5). The early efforts on clinical deidentification began, most notably, with the i2b2 challenges, such as the 2006 and 2014 i2b2 deidentification challenges, which featured a corpus of 1304 medical records from 296 patients with diabetes (6). Similarly, other notable works include the 2016 CEGS N-GRID shared task, which highlights the rich diversity of protected health information categories and demonstrates the substantial difficulty of developing deidentification systems that can generalize across datasets and clinical domains (7). Beyond structured and unstructured text notes, a rapidly growing source of clinical data arises from spoken interactions between doctors and patients, audio recordings of clinical consultations, patient interviews, and telemedicine sessions. These data are captured, stored, and processed for automated clinical documentation, personalized patient care, quality assurance and medical education (8, 9). However, these data

medRxiv preprint doi: https://doi.org/10.64898/2026.04.13.26349911; this version posted April 14, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

introduce unique privacy challenges because, in addition to the textual content of the conversation containing SHI entities, the acoustic signal itself carries personally identifiable information under regulations such as the Health Insurance Portability and Accountability Act (HIPAA) (10, 11) and the European General Data Protection Regulation (12). Recent studies have demonstrated that speaker identification models can reidentify patients under certain adversarial conditions (13). Studies have shown that when private information is removed or deidentified in accordance with privacy regulations, such as HIPAA, using techniques like privacy-aware learning and deidentification methods (6, 14, 15), patients are more willing to share their medical records for study purposes (16).

Additionally, the use of large language models (LLMs) in medical training environments has raised concerns about patient data confidentiality, clinical reasoning and academic integrity (17). Therefore, deidentification is required for data sharing to ensure compliance with regulations and privacy protection. However, for speech-based clinical data, generating deidentified EHRs becomes more complex, as it first involves transcribing the audio signal to text and, secondly, identifying and classifying SHI entities from the transcribed data. During this complex task, misrecognition can degrade SHI recognition performance (18, 19). Furthermore, the multilingual and code-mixed nature of clinical conversations adds additional complexity to both the ASR and NER stages (20-24). The corpora designed by He, Xuehai, et al (25) support dialogue generation models for telemedicine applications. However, this dataset does not include annotations for sensitive health information (SHI).

In contrast, time normalization corpora (2012 i2b2 for clinical text) primarily focused on aligning clinical events within discharge summaries to a temporal timeline (26). To address these concerns, studies have demonstrated that obfuscating speaker attributes while retaining linguistic content can mitigate these risks (27-29). However, studies further revealed that anonymization strategies may have variable effects across different speech disorders,

medRxiv preprint doi: https://doi.org/10.64898/2026.04.13.26349911; this version posted April 14, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

suggesting that domain-specific methods are needed to balance privacy protection with task-relevant utility (30). Furthermore, the annotated corpora and research focused on speech-based deidentification are lacking. Cohn et al. (31) introduced the first benchmark that focuses on audio-based deidentification, integrating real medical records and conversational speech to detect audio segments containing sensitive entity mentions. They also released a new benchmark dataset by annotating the Switchboard (32) and Fisher (33) speech corpora, enabling systematic evaluation of the feasibility and performance of deidentification at the audio level. Therefore, effective clinical speech AI systems must reconcile the imperatives of accurate speech processing with rigorous privacy considerations, highlighting the urgent need for contributions in this area.

To address this gap, building upon the success and insights of the 2023 competition (34, 35), we organized the AI-Cup 2025 Medical Speech Sensitive Personal Data Recognition and Normalization (36), focusing on dual computational challenges: Task 1, medical speech transcription systems must accurately transcribe speech into text; Task 2, the resulting transcriptions must undergo medical speech-sensitive health information (SHI) deidentification to detect and appropriately classify mentions of SHI within the text stream. The challenge provided participants with training data to develop solutions that balance transcription fidelity with robust protection of sensitive health information. This study provides an in-depth overview of the competition and observations on the results presented by participants at the International Workshop on Deidentification of Electronic Medical Record Notes (*IW-DMRN*) on 9th to 13th August 2025, Taipei, Taiwan.

In addition to the interdisciplinary innovative solutions at the intersection of speech processing, privacy-aware learning, and clinical NLP, the major contributions include the benchmark dataset specifically designed for SHI recognition in medical speech (37), thereby advancing speech deidentification technologies. Moreover, the constructed dataset not only helps

medRxiv preprint doi: https://doi.org/10.64898/2026.04.13.26349911; this version posted April 14, 2026. The copyright holder for this preprint **(which was not certified by peer review)** is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

generative models better understand the correlation between sensitive content and acoustic features but also becomes the first study that jointly evaluates ASR and SHI deidentification on real-world medical speech data, thereby also becomes the first work that bridges the gap between text-based deidentification and the practical requirements of processing spoken clinical interactions.

Results

The 2025-SREDH-AICup-SHI-Speech-Corpus Creation

To evaluate the performance of various methods developed by the participating teams in the SREDG/AI CUP 2025 voice deidentification competition, this study released the curated clinical speech dataset (37). Reconstructed and augmented from the OpenDeID v2 corpus developed during SREDH/AI Cup 2023 (34) and the Dataset for Automated Medical Transcription (DAMT) (38). We then extend this by adding audio excerpts from the licensed medical drama content of the Taiwan Public Television Service (PTS) series (39). The final corpus comprises 20 hours of audio recordings, sampled at 16 kHz and in mono. Additionally, a manually verified transcript accompanies each audio file, generated using a hybrid ASR–human correction pipeline.

For the competition SREDH/AI CUP 2025, this study further divided the dataset into a training set (1,539 audio clips), with a total duration of 10 hours, a validation set (775 audio clips), with a total duration of 5 hours, and a testing set (710 audio clips), with a total duration of 5 hours. Overall, the dataset contains 7,830 SHI annotations. Figure 1 illustrates a sample segment of an electronic medical record (EMR) based spoken-style report adapted for recordings. In this example, the highlighted text indicates the SHI entities that annotators should annotate. Table 1 lists all SHI categories and their corresponding subtypes defined in the annotation guidelines, along with representative examples for each.

medRxiv preprint doi: https://doi.org/10.64898/2026.04.13.26349911; this version posted April 14, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

SREDH/AI CUP 2025 Competition

The competition was hosted on Codabench (36), an online platform for benchmarking and evaluation that supports challenges, datasets, and automated evaluation. The competition attracted 246 participants, who formed 149 teams, each led by a designated team leader who played a pivotal role throughout. The participants included individuals from multidisciplinary fields, making this competition more diverse, and the teams used diverse methodologies tailored to their core subject domains. The diverse engagement not only contributed to the diversity of approaches but also demonstrated their commitment and enthusiasm toward the challenge. During the development phase, only three model prediction submissions were allowed per day and evaluated via a public leaderboard. In the final evaluation phase, each team was allowed to submit up to six prediction files over two days. A total of 51 teams successfully submitted prediction results for the final test set, yielding 160 valid submissions, which were evaluated and ranked using the MER and macro F1-score.

The best performance scores achieved for Tasks 1 and 2 were a MER of 0.1147 and a macro F1-score of 0.7103, respectively. On average, across all submissions, MER of 0.3593 and F1-score of 0.2696. These results highlight the challenging nature of the tasks, which required strong analytical skills, effective problem-solving, and diverse modeling strategies. The final rankings were determined by averaging the individual ranks for both tasks for each team and sorting them by the average. Table 3 lists all teams whose Task 1 performance exceeded the baseline score of 0.138, and Table 4 presents all teams whose Task 2 performance exceeded the baseline score of 0.3804.

Beyond the comparative assessment, we conducted a comprehensive analysis to evaluate outcomes and contextual factors. The results highlighted that the distribution for subtask 1 is heavily reliant on Neural and transformer-based architectures (Figure 2). These results also show that the participants used Whisper (a transformer-based model) during their practical

medRxiv preprint doi: https://doi.org/10.64898/2026.04.13.26349911; this version posted April 14, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

deployments. The other notable models used by participants during the practical deployment include WhisperX, Parakeet-tdt, and Conformer. The presence of long-tail diversity in approaches is also evident in the "*Other*" category. The post-competition survey received 65 valid responses, and the results reveal that LLMs were prevalent across both tasks: 97.5% and 100% of the teams adopted LLMs for Task 1 and Task 2, respectively (Figures 2 and 3), showcasing a trend in NLP, where focus is more on the usage of LLMs to address complex speech and text processing tasks.

Whisper Model Evaluation for MER Estimation

To establish a reliable baseline for de-identification, we first evaluated multiple Whisper models on our dataset within the context of the AI CUP competition. Whisper is a multilingual ASR model that utilizes a <|language|> token to determine which specific language-to-text transcription task to execute. The goal was to quantify the Minimal Error Rate (MER) of automatic transcription, which directly impacts downstream de-identification performance. A critical component of the Whisper architecture is Language Identification (LID). Before every transcription, Whisper performs LID to generate a language-probability distribution. It then selects the language with the highest probability to populate the <|language|> token (40). We aim to verify whether Whisper's LID alone can reliably facilitate transcription for specific languages. Table 5 presents the MER across three distinct data distributions from the test set—split into mixed-language and language-independent subsets—where language tokens were automatically generated via LID. Table 5 also lists the MER results obtained by manually specifying the correct language token. It can be observed that the MER decreased slightly across the tiny, small, medium, large, and large-v3 models. Conversely, the MER for the base model showed a marginal increase of 0.018. These results indicate that the LID accuracy is already robust enough to identify Chinese and English speech in our dataset. Table 5 further

medRxiv preprint doi: https://doi.org/10.64898/2026.04.13.26349911; this version posted April 14, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

showcases the performance of Whisper's English-only models. While the tiny, base, and small models exhibited a significant decrease in MER on English corpora, the medium model surprisingly showed a slight increase.

Next, we investigated the extent to which fine-tuning improves Whisper's performance. During the data preprocessing phase, we applied Whisper Text Normalization to the transcriptions to align our data with the stylistic patterns and formatting encountered during the model's pre-training stage. For the fine-tuning process, we employed LoRA for fine-tuning, with hyperparameters set to rank=32, alpha=64, and dropout=0.1 (See detailed specifications in the Supplementary Note 4). Our methodology involved two primary strategies, Bilingual joint training and Bilingual independent training. For both strategies, we applied data augmentation (time stretching and pitch scaling) specifically to the Chinese audio. Table 6 presents the MER results for joint and independent training on the test set. Overall, the top four performing configurations all resulted from the independent training strategy, with whisper-large-v3 achieving the best overall MER of 0.1227.

Fine-tuned LLM performance

We further trained six different language models on the provided dataset, including both general-purpose and medical-domain-specific models. The models include Qwen2.5, Llama3.1, Gemma 2-9B, Mistral, Med42-8B, and OpenBioLLM-8B. All models were fine-tuned on the competition training dataset using a consistent training configuration to ensure fair comparison. This allowed us to assess the effect of model architecture and domain adaptation on de-identification performance. We also included the task-specific instructions to guide the models in identifying and labelling sensitive entities in clinical text. Qwen 2.5 - 7B exhibited a balanced performance with a precision of 0.8871 and a recall of 0.8372, leading to an F1 score of 0.8614, indicating a strong ability to correctly identify and comprehensively capture

medRxiv preprint doi: https://doi.org/10.64898/2026.04.13.26349911; this version posted April 14, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

sensitive entities within clinical text while effectively minimising both false positives and missed entities. Llama 3.1 - 8B slightly outperformed Qwen 2.5 with a precision of 0.8871 and a recall of 0.8470, resulting in an F1 score of 0.8666. Mistral - 7B also showed competitive performance, achieving a precision of 0.8852, a recall of 0.8391, and an F1 score of 0.8615, suggesting effectiveness in identifying a wide range of sensitive health information entities. Med42 - 8B achieved a precision of 0.8893 and a recall of 0.8329, resulting in an F1 score of 0.8602. Its performance remained competitive with the general-purpose models; it did not surpass them. Gemma 2 - 9B exhibited a different performance pattern, with a precision of 0.8572 and a significantly lower recall of 0.7462, resulting in an F1 score of 0.7979, suggesting that the model produced fewer false positives; it struggled to capture a large proportion of sensitive entities, indicating limited coverage of the target annotations. OpenBioLLM-8 B performed well, achieving a precision of 0.8863 and a recall of 0.8299, resulting in an F1 score of 0.8572, demonstrating its potential for clinical de-identification tasks. Supplementary Table 14 presents detailed scores for each entity type across models, highlighting differences in model performance on DOCTOR, DATE, PATIENT, and other SHI categories. Overall, these findings indicate that both general-purpose large language models and domain-specific models can achieve strong performance on clinical de-identification tasks.

Furthermore, Analysing the results per-type F1 carefully, the winning strategy is that different models are specialists at different entity types, like for example Gemma dominates PROFESSION/ROOM, Qwen and LLaMa together are good at PHONE/ZIP, Med42 performs well in DURATION/HOSPITAL, and OpenBioLLM performs well on PERSONALNAME/ORGANIZATION, etc. Therefore, based on the mixture of experts (MoE), the key principle is sparse gating, i.e. activate only the best expert per input. Based on these conditions we implemented a type-level sparse gate inspired by the MoE (41, 42) routing mechanism in transformer architectures. This design ensures that for common types, consensus

medRxiv preprint doi: https://doi.org/10.64898/2026.04.13.26349911; this version posted April 14, 2026. The copyright holder for this preprint **(which was not certified by peer review)** is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

among strong experts is required before accepting a prediction, while for rare types where all models score near-zero, even a single expert's prediction is admitted maximizing the recall. We achieved a Macro F1 of 0.6965 and Micro F1 of 0.8629, representing a +0.041 Macro improvement over the best individual model (Med42: 0.6559) and a +0.002 Micro improvement over the best individual Micro score (LLaMA: 0.8666) (supplementary note 6).

Discussion

Overall, this work underscores the key insights from the AI-Cup 2025 competition, demonstrating participant's innovative approaches to medical speech processing. The widespread use of LLMs, particularly in ASR and SHI recognition tasks, underscores their growing impact on medical artificial intelligence. As LLM technologies continue to advance, they are expected to drive more robust, scalable solutions to future healthcare challenges. Furthermore, clinical and real-world applications increasingly rely on speech; these findings demonstrate that audio de-identification is not a straightforward extension of text-based approaches but a distinct problem requiring dedicated methods, metrics, and security/threat analysis models. Hence, there is a strong need to establish standardized benchmarks and evaluation frameworks to enable responsible data sharing and deployment.

For Task 1, the most used architecture for ASR was OpenAI's Whisper (43). For Task 2, Qwen3 (44) was the most frequently adopted model, followed by DeepSeek (45) and Llama 3 (46). This distribution provides a clear overview of participants' architectural preferences and highlights the growing dominance of LLM-based solutions in clinical NLP applications.

Based on the results and the survey data, the top-performing teams used diverse and innovative approaches across both subtasks. *TEAM_7526* (47), who ranked first, instead of general transcription-based pipelines, adopted a dual-modality approach using CriperWhisper (48) and FireRedASR (49) for transcription, aiming to suppress automatic speech recognition while preserving human intelligibility. They employed Qwen3 (44) for SHI recognition. In contrast,

medRxiv preprint doi: https://doi.org/10.64898/2026.04.13.26349911; this version posted April 14, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

*TEAM_*7882 (50), the second-ranked approach, relied on an upstream ASR model based on Whisper-large-v3-zh (51, 52), a model specifically optimised for Mandarin ASR, and conducted comparative experiments between decoder-layer fine-tuning and full-parameter tuning. At the same time, the *TEAM_7913* (53), in third place, employed a dual-modality design using Parakeet (54) and Whisper (43) for multilingual transcription and DeBERTa v3(55)/mDeBERTa v3 (56) for SHI entity recognition.

Beyond the top-ranking submission, the remaining high-performing systems adopted variations of transcription-driven architectures, differing primarily in how they handle data, ASR uncertainty, alignment granularity, and post-processing strategies. While their results did not meet the privacy–utility balance of the top-ranked teams, their results underscore the viability of ASR-based pipelines under appropriate design constraints. For example, several teams (57-59) reformulated the training data using instructional-style formatting to preserve contextual semantics better, enabling the models to understand labelling intent better. LLMs were also employed to paraphrase a subset of training samples for semantic augmentation. Some teams (57, 58, 60) conducted comparative studies on coarse-grained entity grouping and found that applying LoRA (61) fine-tuning across six broader categories improved generalisation. In addition, several teams leveraged retrieval-augmented generation, a core technique highlighted in the AI-Cup 2023 challenge (34), to construct vector-based similarity indices to dynamically select semantically relevant examples for few-shot prompting. This adaptive approach demonstrated better accuracy than conventional fixed-shot methods, especially in complex or ambiguous input scenarios. For detailed results, check the comprehensive results for the top 10 teams in the Supplementary Tables 1-10.

The competition produced different high-performing solutions using different strategies and methods. To investigate whether task-specific adaptation could further improve performance beyond the competition submissions, we fine-tuned six large language models on the training

medRxiv preprint doi: https://doi.org/10.64898/2026.04.13.26349911; this version posted April 14, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

dataset to establish a stronger benchmark system through more detailed analysis. Keeping in view the current issues in the domain, like 1) Class imbalance problem: Existing models achieve ~80-90% on common entities (names, dates), but most of the time these models achieve low results on rare entities (profession, duration). 2) Conversation vs Clinical Notes: studies mostly focus on structured clinical documentation, not natural doctor-patient dialogues. To address these issues, we used a controlled/unified training and evaluation pipeline to ensure a fair comparison across six models both general purpose and medical-domain models.

Methods

The 2025-SREDH-AICup-SHI-Speech-Corpus

SpeechDeID was constructed from multiple medical-domain datasets from national and international sources, including the OpenDeID v2 corpus, the DAMT, and a dataset from the Taiwan PTS. During construction, we followed the HSA Study SHI Corpus–Annotation Guidelines (62), with details provided in the supplementary note 1.

The dataset preparation workflow involved several key steps. First, selected EMR documents from the OpenDeID v2 corpus were rewritten by two domain specialists into conversational, spoken-style text, adhering to SHI annotation categories. Each rewritten script contained approximately 100–200 words. Next, trained personnel recorded the spoken versions of these scripts. A total of 25 participants from the College of Foreign Languages at the National Kaohsiung University of Science and Technology (9 male and 16 female) recorded approximately 10–20 audio samples each. Participants received instructions on speech rate, pause placement, and pronunciation to ensure consistency and had time to rehearse before recording. Recordings lasted 1.0-1.8 minutes and underwent quality checks before further processing.

Before being incorporated into the dataset, all subtitle files and speech recordings went through two additional processing stages. The first stage involved aligning each audio file with its

medRxiv preprint doi: https://doi.org/10.64898/2026.04.13.26349911; this version posted April 14, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

corresponding subtitle file to generate word-level timestamp transcriptions using the Montreal Forced Aligner (63). The second stage applied voice activity detection (VAD) (64) to segment long-form audio, retaining only human speech. To ensure that no audio exceeded 30 seconds, concatenation and truncation were applied to the resulting speech segments. Aligned subtitle timestamps were then mapped to their corresponding audio segments and manually verified.

For data sources not originating from the OpenDeID v2 corpus, such as the DAMT and the PTS dataset, the process began at the VAD stage, as these sources already contained speech data.

The recorded data, adapted from the OpenDeID v2 corpus, were produced in three distinct phases to ensure no overlap with the original reports. In Phase 1, recorded audio was evenly divided between the training and validation sets. Phases 2 and 3 provided additional audio exclusively for the test set to maintain independence and prevent training on the same reports. The speech data from the DAMT were randomly split into training and validation sets, with no constraints on the underlying reports. At the same time, the PTS audio, characterised by heterogeneity and imbalanced durations, was divided evenly into training and testing halves.

The final competition test set contained only audio recorded by our team and PTS-sourced speech, excluding any publicly available online data. This design ensured that participants could not reconstruct answers from publicly available sources, thereby maintaining fairness and integrity in the evaluation.

In summary, the dataset preparation workflow involved several steps. First, the collected EMR documents were rewritten into more conversational, spoken-style text. Second, trained personnel recorded the spoken versions of the rewritten reports. Third, voice activity detection was applied to remove silent segments from the recordings. Finally, the resulting speech segments were concatenated and truncated to ensure that each audio clip was no longer than 30 seconds. For data sources not originating from the OpenDeID v2 corpus, the process began

medRxiv preprint doi: https://doi.org/10.64898/2026.04.13.26349911; this version posted April 14, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

directly from the third step. Figure 4 illustrates a simplified workflow of the dataset creation pipeline. Furthermore, the details about the prediction files are provided in Table 2, guiding participants on preparing complaint submissions and understanding the evaluation criteria, with a focus on entity type and timestamp accuracy.

Competition design

The competition consists of two subtasks, both based on the same set of medical speech audio recordings. Task 1 focuses on generating audio transcriptions to encourage the development of speech recognition models tailored for clinical environments. Task 2 involves identifying SHI within the transcribed text. In addition to labelling the entities embedded in the speech, participants must provide the corresponding start and end timestamps for each entity. The submissions are encouraged not only to reduce transcription error rates but also to handle information associated with specific temporal segments of the audio. Figure 5 shows the overview of the AI-Cup 2025 competition task structure. The competition comprises two sequential tasks for processing medical audio recordings. Figure 6 shows an illustrative example to help the participating teams better understand the evaluation methodology. In the left example, both the predicted and ground-truth labels are DOCTOR; thus, the overlapping portion is counted as TP, whereas the extra predicted segment is counted as FP. In the middle example, the predicted label is PATIENT, whereas the ground-truth label is DOCTOR. Because the prediction is incorrect, the entire ground-truth segment is counted as FN, and the predicted segment is counted as FP. Finally, precision, recall, and F1-score for each category are calculated based on their accumulated TP, FP, and FN values.

These duration-based definitions allow for fine-grained temporal alignment scoring, which is especially relevant when SHI spans partially overlap or slightly differ in boundaries. Final micro and macro F1-scores are computed by aggregating across all SHI types. The ranking

medRxiv preprint doi: https://doi.org/10.64898/2026.04.13.26349911; this version posted April 14, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

procedure of this competition follows the AI-Cup 2023 rules (34, 35). The rankings for Tasks 1 and 2 are computed separately, and their ranks are then summed to determine the final overall ranking. Details of evaluation metrics and criteria are provided in the supplementary note 2.

Fine tuning of LLMs

The preprocessing pipeline transformed the raw annotation files into an instruction-tuning format (Supplementary Note 3) suitable for generative models. The raw data in the transcript file contains textual transcripts of medical conversations, where each entry is a pair: a unique file identifier and the corresponding transcript. The annotation file contains sensitive health information (SHI) annotations, including identifiers, entity types, timestamps, and annotated entity text. The pipeline then performs dataset parsing, annotation alignment, dataset merging, and conversion into a structured instruction-response format. These two files together provide the textual input and the corresponding entity annotations required for training. There are two separate training phases; therefore, we merged the transcripts and annotations from both phases to produce a consolidated dataset containing all transcripts and their corresponding SHI annotations.

Furthermore, to enable large language models to extract entities by generating structured output, we reformulated the named-entity recognition task as an instruction-following generation task rather than a token-classification task. More details on the Instruction following generation are presented in *Supplementary Notes 3 and 4*. Additionally, this preprocessing pipeline also preserves entity timestamps from the original annotations. These timestamps are stored alongside the instruction examples.

To ensure a comprehensive evaluation, we fine-tuned both domain-specific medical language models and general-purpose instruction-tuned models to assess their effectiveness for SHI extraction. This comparison also allows us to examine whether models trained on biomedical

medRxiv preprint doi: https://doi.org/10.64898/2026.04.13.26349911; this version posted April 14, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

data provide advantages over strong general-purpose LLMs. Models were selected to represent a range of architectures, parameter sizes, and training domains, including both general-purpose and medical-domain models (Supplementary Table 11). The model receives the instruction and is trained to generate corresponding entity annotations (Supplementary Note 3). During training, each example was converted into a formatted conversational prompt according to the models chat template. The formatted prompt was then tokenised using the corresponding model tokenizer with a maximum sequence length of 768, and padding was dynamically applied with truncation enabled. Labels were constructed using the causal language modelling objective, where the model learns to predict the next token in the assistant response. More details on the Alpaca-style instruction format, timestamp alignment, and the instruction-based NER formulation are provided in *Supplementary Note 3*, *Supplementary Note 5*, *Supplementary Note 6*, respectively. Finally, we implemented a unified training framework applying 4-bit quantisation and LoRA adaptation across all architectures. Each model was loaded, quantised, injected with LoRA adapters, and formatted using its appropriate chat template. Checkpoints were evaluated periodically, retaining the best-performing checkpoint for downstream tasks, including model comparison and ensemble methods. Full training configuration and LoRA design details are provided in Supplementary Tables 12 and 13, respectively.

Ethics approval

The Human Research Ethics Committee of The University of New South Wales (UNSW Sydney) gave ethical approval for this work (Approval No. HC17749). The datasets had been de-identified prior to access and use by the authors. The corpus used in the SREDH/AICUP 2025 competition is publicly accessible and is subject to approval by signing a data-usage agreement. Access can be requested through the SREDH consortium git profile here (https://github.com/SREDH-Consortium).

medRxiv preprint doi: https://doi.org/10.64898/2026.04.13.26349911; this version posted April 14, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

Code availability

More details about the competition and the baseline code can be found here https://www.codabench.org/competitions/4890/?secret_key=38d92718-cc4d-4907-9c65-c73419671268.

Acknowledgments

This study was funded by the SREDH/AICUP 2025, which occurred under the auspices of the Competition Guidance Unit within the Department of Information, Technology, and Education at the Ministry of Education (MoE), Taiwan. The Competition Planning Unit of the Artificial Intelligence and Annotation Data Collection Program Office, MoE, Taiwan provided strategic coordination and oversight. This competition is a collaborative effort involving the Intelligent Systems Laboratory (ISLab) and the Wireless Networking and Mobile Computing Research Lab within the Department of Electrical Engineering, as well as the Department of Applied English at the National Kaohsiung University of Science and Technology; the Machine Learning and Bioinformatics Lab within the Department of Computer Science at the University of Taipei; the Department of Bioinformatics and Medical Engineering at Asia University; and the SREDH Consortium. JJ was funded by the Australian National Health and Medical Research Council (Grant GNT1192469). We acknowledge the funding support received through the Research Technology Services at the University of New South Wales Sydney and ASUSTeK Computer Inc and the National Science and Technology Council under the grant number NSTC 114-2637-8-992-007-. We also thank the Secure Research Environment for Digital Health (SREDH) Consortium Translational Cancer Bioinformatics Working Group for providing access to the OpenDeID corpus.

Competing interests

medRxiv preprint doi: https://doi.org/10.64898/2026.04.13.26349911; this version posted April 14, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

The authors declare no competing interests or conflicts of interest regarding the content or findings presented in this study.

References

1. Coorevits P, Sundgren M, Klein GO, Bahr A, Claerhout B, Daniel C, et al. Electronic health records: new opportunities for clinical research. J Intern Med. 2013;274(6):547-60.
2. Colquhoun DA, Shanks AM, Kapeles SR, Shah NR, Saager L, Vaughn MT, et al. Considerations for Integration of Perioperative Electronic Health Records Across Institutions for Research and Quality Improvement: The Approach Taken by the Multicenter Perioperative Outcomes Group. Anesth Analg. 2020;130(5):1133-46.
3. Romanowski B, Ben Abacha A, Fan YD. Extracting social determinants of health from clinical note text with classification and sequence-to-sequence approaches. J Am Med Inform Assn. 2023;30(8):1448-55.
4. Birkhead GS, Klompas M, Shah NR. Uses of Electronic Health Records for Public Health Surveillance to Advance Public Health. Annu Rev Publ Health. 2015;36:345-59.
5. Hillestad R, Bigelow J, Bower A, Girosi F, Meili R, Scoville R, et al. Can electronic medical record systems transform health care? Potential health benefits, savings, and costs. Health Aff (Millwood). 2005;24(5):1103-17.
6. Stubbs A, Kotfila C, Uzuner Ö. Automated systems for the de-identification of longitudinal clinical narratives: Overview of 2014 i2b2/UTHealth shared task Track 1. J Biomed Inform. 2015;58:S11-S9.
7. Stubbs A, Filannino M, Uzuner Ö. De-identification of psychiatric intake records: Overview of 2016 CEGS N-GRID shared tasks Track 1. J Biomed Inform. 2017;75:S4-S18.
8. Quiroz JC, Laranjo L, Kocaballi AB, Berkovsky S, Rezazadegan D, Coiera E. Challenges of developing a digital scribe to reduce clinical documentation burden. Npj Digit Med. 2019;2.
9. van Buchem MM, Boosman H, Bauer MP, Kant IMJ, Cammel SA, Steyerberg EW. The digital scribe in clinical practice: a scoping review and research agenda. Npj Digit Med. 2021;4(1).
10. Jonnagaddala J, Wong ZS-Y. Privacy preserving strategies for electronic health records in the era of large language models. Npj Digit Med. 2025;8(1):34.
11. Services USDoHaH. Health Insurance Portability and Accountability Act of 1996. In: HHS, editor. Washington, DC: ASPE.
12. Union E. Regulation (EU) 2016/679 of the European Parliament and of the Council on the protection of natural persons with regard to the processing of personal data and on the free movement of such data (General Data Protection Regulation). 4 May 2016. p. 1 – 88.
13. Wiepert D, Malin BA, Duffy JR, Utianski RL, Stricker JL, Jones D, et al. Reidentification of Participants in Shared Clinical Data Sets: Experimental Study. Jmir Ai. 2024;3.
14. Uzuner Ö, Luo Y, Szolovits P. Evaluating the state-of-the-art in automatic de-identification. J Am Med Inform Assn. 2007;14(5):550-63.
15. Wang CK, Wang FD, Lee YQ, Chen PT, Wang BH, Su CH, et al. Principle-Based Approach for the De-Identification of Code-Mixed Electronic Health Records. Ieee Access. 2022;10:22875-85.
16. King T, Brankovic L, Gillard P. Perspectives of Australian adults about protecting the privacy of their health information in statistical databases. Int J Med Inform. 2012;81(4):279-89.
17. Tran M, Balasooriya C, Jonnagaddala J, Leung GKK, Mahboobani N, Ramani S, et al. Situating governance and regulatory concerns for generative artificial intelligence and large language models in medical education. Npj Digit Med. 2025;8(1).
18. Mani A, Palaskar S, Meripo NV, Konam S, Metze F, editors. ASR Error Correction and Domain Adaptation Using Machine Translation. ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP); 2020 4-8 May 2020.
19. Gaur Y, Lasecki WS, Metze F, Bigham JP. The effects of automatic speech recognition quality on human transcription latency. Proceedings of the 13th International Web for All Conference; Montreal, Canada: Association for Computing Machinery; 2016. p. Article 23.

medRxiv preprint doi: https://doi.org/10.64898/2026.04.13.26349911; this version posted April 14, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

20. Hou S-Y, Wu Y-L, Chen K-C, Chang T-A, Hsu Y-M, Chuang S-J, et al. Code-Switching Automatic Speech Recognition for Nursing Record Documentation: System Development and Evaluation. JMIR Nursing. 2022;5(1):e37562.
21. Barrault L, Chung Y-A, Meglioli MC, Dale D, Dong N, Duquenne P-A, et al. Joint speech and text machine translation for up to 100 languages. Nature. 2025;637(8046):587-93.
22. Meeus Q, Moens MF, editors. Msner: A multilingual speech dataset for named entity recognition. Proceedings of the 20th Joint ACL-ISO Workshop on Interoperable Semantic Annotation@ LREC-COLING 2024; 2024.
23. Dowlagar S, Mamidi R. A code-mixed task-oriented dialog dataset for medical domain. Computer Speech & Language. 2023;78:101449.
24. Tian J, Yu J, Zhang C, Weng C, Zou Y, Yu D. Lae: Language-aware encoder for monolingual and multilingual asr. arXiv preprint arXiv:220602093. 2022.
25. He X, Chen S, Ju Z, Dong X, Fang H, Wang S, et al. Meddialog: Two large-scale medical dialogue datasets. arXiv preprint arXiv:200403329. 2020.
26. Sun W, Rumshisky A, Uzuner O. Evaluating temporal relations in clinical text: 2012 i2b2 Challenge. J Am Med Inform Assn. 2013;20(5):806-13.
27. Qian J, Han F, Hou J, Zhang C, Wang Y, Li X-Y, editors. Towards Privacy-Preserving Speech Data Publishing. INFOCOM; 2018.
28. Tran M, Soleymani M. Privacy-preserving Representation Learning for Speech Understanding. arXiv preprint arXiv:231017194. 2023.
29. Perero-Codosero JM, Espinoza-Cuadros FM, Hernández-Gómez LA. X-vector anonymization using autoencoders and adversarial training for preserving speech privacy. Computer Speech & Language. 2022;74:101351.
30. Tayebi Arasteh S, Arias-Vergara T, Pérez-Toro PA, Weise T, Packhäuser K, Schuster M, et al. Addressing challenges in speaker anonymization to maintain utility while ensuring privacy of pathological speech. Communications Medicine. 2024;4(1):182.
31. Cohn I, Laish I, Beryozkin G, Li G, Shafran I, Szpektor I, et al., editors. Audio De-identification - a New Entity Recognition Task2019 June; Minneapolis, Minnesota: Association for Computational Linguistics.
32. Godfrey JJ, Holliman EC, McDaniel J, editors. SWITCHBOARD: telephone speech corpus for research and development. [Proceedings] ICASSP-92: 1992 IEEE International Conference on Acoustics, Speech, and Signal Processing; 1992 23-26 March 1992.
33. Cieri C, Miller D, Walker K, editors. The Fisher Corpus: a Resource for the Next Generations of Speech-to-Text2004 May; Lisbon, Portugal: European Language Resources Association (ELRA).
34. Dai H-J, Mir TH, Chen C-T, Chen C-C, Yang H-P, Lee C-H, et al. Leveraging large language models for the deidentification and temporal normalization of sensitive health information in electronic health records. Npj Digit Med. 2025;8(1):517.
35. Mir TH, Yang H-P, Chou Y-Y, Teng Y-C, Liao W-H, Lin Y-C, et al., editors. Deidentification and Temporal Normalization of the Electronic Health Record Notes Using Large Language Models: The 2023 SREDH/AI-Cup Competition for Deidentification of Sensitive Health Information2025; Singapore: Springer Nature Singapore.
36. Codabench. AICUP Medical Voice Sensitive Personal Data Recognition Competition: Codabench; 2025 [Available from: https://www.codabench.org/competitions/4890/?secret_key=38d92718-cc4d-4907-9c65-c73419671268.
37. Dai H-J, Fang L-C, Mir TH, Chen C-T, Feng H-H, Lai J-R, et al. Corpus for Benchmarking Clinical Speech De-identification. 2026:2026.03.31.26349906.
38. Kazi N KM, Kanewala U, Kahanda I. Dataset for Automated Medical Transcription. In: Zenodo, editor. 2020.
39. Taiwan PTS. Content licensed for academic research dataset construction. Mad Doctor, Wake up, Coordinator, A Young Doctor in the Big Hospital. Taiwan: Public Television Service (Taiwan); n.d. p. https://www.pts.org.tw/.
40. Peng P, Yan B, Watanabe S, Harwath D, editors. Prompting the hidden talent of web-scale speech models for zero-shot task generalization. Proc Interspeech 2023; 2023.

medRxiv preprint doi: https://doi.org/10.64898/2026.04.13.26349911; this version posted April 14, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

41. Pal A, Sankarasubbu M. OpenBioLLMs: Advancing Open-Source Large Language Models for Healthcare and Life Sciences: Hugging Face; 2024 [Available from: https://huggingface.co/aaditya/OpenBioLLM-Llama3-70B.
42. Christophe C, Kanithi PK, Raha T, Khan S, Pimentel MA. Med42-v2: A suite of clinical llms. arXiv preprint arXiv:240806142. 2024.
43. Radford A, Kim JW, Xu T, Brockman G, McLeavey C, Sutskever I, editors. Robust speech recognition via large-scale weak supervision. International conference on machine learning; 2023: PMLR.
44. Yang A, Li A, Yang B, Zhang B, Hui B, Zheng B, et al. Qwen3 technical report. arXiv preprint arXiv:250509388. 2025.
45. Guo D, Yang D, Zhang H, Song J, Wang P, Zhu Q, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. arXiv preprint arXiv:250112948. 2025.
46. Grattafiori A, Dubey A, Jauhri A, Pandey A, Kadian A, Al-Dahle A, et al. The llama 3 herd of models. arXiv preprint arXiv:240721783. 2024.
47. Huang C-LaH-JD. Temporal Subword De-identification of Medical Speech for Privacy Protection Leveraging ASR and LLMs. in . 20th World Congress on Medical and Health Informatics; Taipei, Taiwan.: MedInfo 2025; 2025.
48. Wagner L, Thallinger B, Zusag M. Crisperwhisper: Accurate timestamps on verbatim speech transcriptions. arXiv preprint arXiv:240816589. 2024.
49. Xu K-T, Xie F-L, Tang X, Hu Y. Fireredasr: Open-source industrial-grade mandarin speech recognition models from encoder-decoder to llm integration. arXiv preprint arXiv:250114350. 2025.
50. Hsu Y-CaH-JD. Speech De-identification of Chinese, English and Min-nan: Effectiveness of Chinese-based LLM Model and ASR. 20th World Congress on Medical and Health Informatics; 2025. Taipei, Taiwan: MedInfo 2025; 2025.
51. BELLEGroup. BELLE: Be Everyone’ s Large Language model Engine Online: GitHub; 2023 [Available from: https://github.com/LianjiaTech/BELLE.
52. Ji Y, Deng Y, Gong Y, Peng Y, Niu Q, Ma B, et al. Belle: Be everyone’ s large language model engine. 2023.
53. Chen Y-J, T.-Y. Chang, and T.-H. Chang. . Multistage Method for Named-Entity Recognition in Medical Audio Data. 20th World Congress on Medical and Health Informatics; Taipei, Taiwan.: MedInfo 2025; 2025.
54. NVIDIA. Parakeet-TDT-0.6B-v2 model card Online: Hugging Face; 2025 [Available from: https://huggingface.co/nvidia/parakeet-tdt-0.6b-v2.
55. He P, Gao J, Chen W. Debertav3: Improving deberta using electra-style pre-training with gradient-disentangled embedding sharing. arXiv preprint arXiv:211109543. 2021.
56. He P, Liu X, Gao J, Chen W. Deberta: Decoding-enhanced bert with disentangled attention. arXiv preprint arXiv:200603654. 2020.
57. Jing Jin, Jui-Chien Tsou, Chih-Ho Chen, Po-Jen Ko, Huang T-H. In Proceedings of the 2025 International Workshop on MedInfo.
58. Bo-Han Feng, Yu-Chi Cheng, Wen-Po Lin, You-Hsuan Chang, Vu Thinh Doan, Hsieh K-P. In Proceedings of the 2025 International Workshop on MedInfo.
59. Zheng-Hao L, Dai H-J. In Proceedings of the 2025 International Workshop on MedInfo.
60. Ping-Hsien Lin, Pratham Nandy, Wen C-Y. In Proceedings of the 2025 International Workshop on MedInfo.
61. Hu EJ, Shen Y, Wallis P, Allen-Zhu Z, Li Y, Wang S, et al. Lora: Low-rank adaptation of large language models. Iclr. 2022;1(2):3.
62. Quinn CM, Porwal M, Meagher NS, Hettiaratchi A, Power C, Jonnaggadala J, et al. Moving with the Times: The Health Science Alliance (HSA) Biobank, Pathway to Sustainability. Biomarker Insights. 2021;16:11772719211005745.
63. McAuliffe M, Socolof M, Mihuc S, Wagner M, Sonderegger M, editors. Montreal forced aligner: Trainable text-speech alignment using kaldi. Interspeech; 2017.
64. Tan Z-H, Dehak N. rVAD: An unsupervised segment-based robust voice activity detection method. Computer speech & language. 2020;59:1-21.

medRxiv preprint doi: https://doi.org/10.64898/2026.04.13.26349911; this version posted April 14, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

Figure Legends

Figure 1. Example audio transcript with sensitive health information. This figure illustrates a segment of a deidentified report from the released dataset. All SHI mentions are highlighted in bold and have been replaced with surrogates to preserve privacy. All entity information is entirely fabricated and does not represent real individuals.

Figure 2. The results of model usage distribution for Subtask 1. The figure shows that Subtask 1's distribution relies heavily on Neural and Transformer architectures. The results also show that the Whisper, a Transformer-based model, is widely used in practical deployments. Furthermore, other notable practical deployments include Transformer-based models such as WhisperX, Parakeet-tdt, and Conformer. The "Other" category of models also indicates the presence of long-tail diversity in approaches.

Figure 3. The results of model usage distribution for Subtask 2. Radial schematic illustrating the taxonomy of neural architectures and selected example models. Concentric layers represent increasing specificity, progressing from the central category Neural to Transformers, then to architectural subclasses, and finally to individual systems positioned on the outer ring (including Qwen-3, Deepseek, Llama-3, Lallama 3.2-instruct, GPT-4, Mixtral, GPT-3.5, Gemini, Gemma-3).

Figure 4. Figure illustrates the overall Pipeline for speech segmentation and alignment used in dataset construction. Schematic overview of the preprocessing workflow. Input audio recordings are first processed using voice activity detection (VAD) to extract speech segments, while corresponding transcription files are simultaneously analyzed with forced alignment (MFA) to obtain word-level timestamps. The outputs from both stages are subsequently integrated to generate the final curated dataset used for downstream analyses.

Figure 5. Overview of the AI-Cup 2025 competition task structure. The competition comprises two sequential tasks for processing medical audio recordings. In Task 1, participants transcribe spoken clinical narratives from audio inputs into free-text format, generating structured text files (task1_answer.txt) containing patient demographic and medical information. In Task 2, participants perform named entity recognition (NER) on the transcribed

medRxiv preprint doi: https://doi.org/10.64898/2026.04.13.26349911; this version posted April 14, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

outputs, identifying and classifying SHI entities, with their corresponding character-level start and end offsets. The output (task2_answer.txt) provides entity-level annotations specifying the filename, entity type, positional offsets, and extracted content for each recognized SHI instance. Together, the two tasks form an end-to-end pipeline for automatic de-identification of medical speech data. All entity information is entirely fabricated and does not represent real individuals.

Figure 6. **Illustrative example of the evaluation framework used for performance assessment.** Diagram demonstrates how predicted outputs are compared with ground-truth annotations to compute evaluation metrics. Labeled spans indicate reference entities (top) and model predictions (below), with color-coded regions denoting true positives, false positives and false negatives. This example is provided to clarify the scoring procedure and assist participating teams in interpreting the evaluation methodology.

Table Legends

Table 1. Table 1 lists all SHI categories and their corresponding subtypes defined in the annotation guidelines, along with representative examples.

Table 2. Table 2 shows an example of the final Task 1 transcript answer format.

Table 3. Lists of teams whose Task 1 performance exceeded the baseline score.

Table 4. List of all teams whose Task 2 performance exceeded the baseline score of 0.3804.

Table 5. The table reports Match Error Rate (MER) for different Whisper model sizes, from Tiny to Large-v3-turbo. MER is shown overall, for English, and for Chinese, with lower values indicating better transcription accuracy. “Language Detect” refers to models that automatically identify the input language, “Specify Language Token” refers to models provided with the target language token, and “English-only” refers to models trained solely on English data.

Table 6. This table presents the performance of Whisper models across different sizes under two bilingual training strategies: joint training and independent training. Performance is reported in terms of Overall MER, English MER, and Chinese MER, allowing direct comparison of how model size and training approach affect recognition accuracy in both languages.

medRxiv preprint doi: https://doi.org/10.64898/2026.04.13.26349911; this version posted April 14, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

Table 7. The comparative performance of the models on Task 2.

Figures

Figure 1. Example audio transcript with sensitive health information. All SHI mentions highlighted have been replaced with surrogates to preserve privacy.

The 57-year-old patient, Ken Moll, identified by ID number 62S021442H, resides on Yale Street, in the city of Andergrove, Tasmania (ZIP Code 2042).
The case is documented under Medical Record 6270214.MFH, with laboratory investigations recorded as Lab No. 62S02144.

Figure 2. The results of model usage distribution for Subtask 1.

Figure 3. The results of model usage distribution for Subtask 2.

medRxiv preprint doi: https://doi.org/10.64898/2026.04.13.26349911; this version posted April 14, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

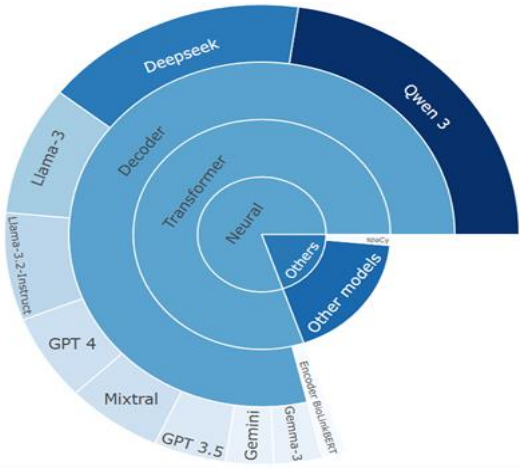


Figure 4. Figure illustrates the overall Pipeline for speech segmentation and alignment used in dataset construction.

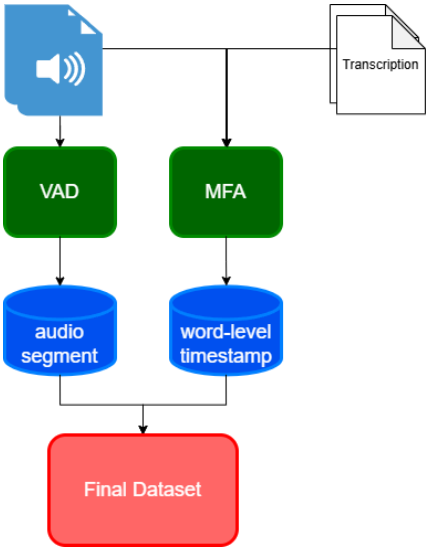


Figure 5. Overview of the AI-Cup 2025 competition task structure. **All SHI mentions have been replaced with surrogates to preserve privacy.**

medRxiv preprint doi: https://doi.org/10.64898/2026.04.13.26349911; this version posted April 14, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

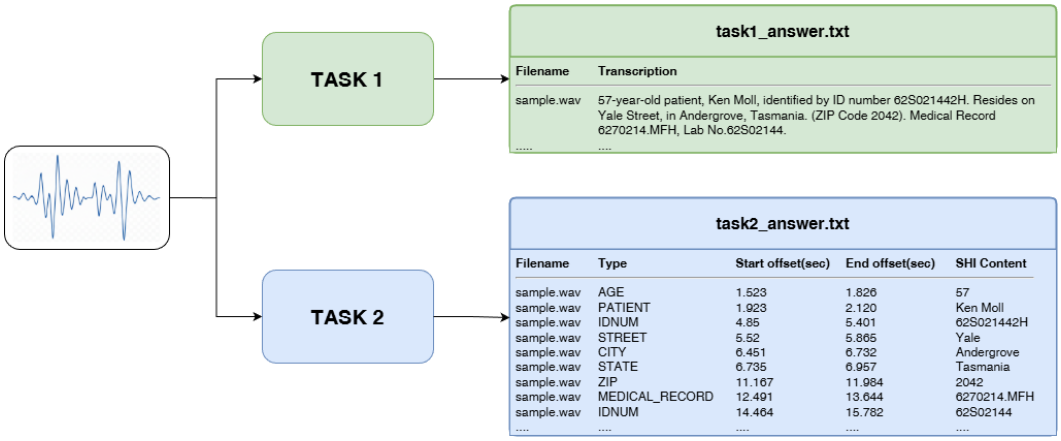
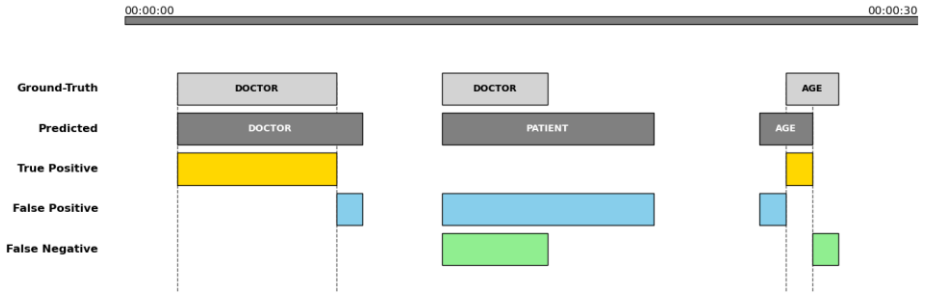

TASK 1
TASK 2
task1_answer.txt
Filename
Transcription
sample.wav
57-year-old patient, Ken Moll, identified by ID number 62S021442H. Resides on Yale Street, in Andergrove, Tasmania. (ZIP Code 2042). Medical Record 6270214.MFH, Lab No.62S02144.
task2_answer.txt
Filename
Type
Start offset(sec)
End offset(sec)
SHI Content
sample.wav AGE 1.523 1.826 57
sample.wav PATIENT 1.923 2.120 Ken Moll
sample.wav IDNUM 4.85 5.401 62S021442H
sample.wav STREET 5.52 5.865 Yale
sample.wav CITY 6.451 6.732 Andergrove
sample.wav STATE 6.735 6.957 Tasmania
sample.wav ZIP 11.167 11.984 2042
sample.wav MEDICAL_RECORD 12.491 13.644 6270214.MFH
sample.wav IDNUM 14.464 15.782 62S02144

Figure 6. Illustrative example of the evaluation framework used for performance assessment.



Tables

Table 1. Lists all SHI categories and their corresponding subtypes defined in the annotation guidelines, along with representative examples.

SHI category	Subcategory	Example
NAME	PATIENT, DOCTOR, USERNAME, PERSONALNAME, FAMILYNAME	Gudara,Pineau, Shaunda, Hinck
PROFESSION	(none)	Lawyer, teacher
LOCATION	ROOM, DEPARTMENT, HOSPITAL, ORGANIZATION, STREET, CITY, STATE, COUNTRY, ZIP, OTHER	PERIOPERATIVE UNIT-POW, MACQUARIE WARD - RHW
AGE	(none)	23, 98
DATE	DATE, TIME, DURATION, SET	September, 26th
CONTACT	PHONE, FAX, EMAIL, URL, IPADDRESS	abc@gmail.com, 194.223.1.1
IDs	SOCIAL_SECURITY_NUMBER , MEDICAL_RECORD_NUMBER , HEALTH_PLAN_NUMBER , ACCOUNT_NUMBER , LICENSE_NUMBER , VEHICLE_ID,DEVICE_ID , BIOMETRIC_ID , ID_NUMBER	MRN: 9174338 IDNUMBER: **12R1500257**
OTHER	(none)	Finger Print, Photograph, Company Logo

medRxiv preprint doi: https://doi.org/10.64898/2026.04.13.26349911; this version posted April 14, 2026. The copyright holder for this preprint **(which was not certified by peer review)** is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

Table 2. Table 2 shows an example of the final Task 1 transcript answer format.

Filename	Type	Start offset (s)	End offset (s)	SHI content (Optional)
Medical_audio_01	**AGE**	1.523	1.826	57
Medical_audio_01	**PATIENT**	1.923	2.120	Ken Moll
Medical_audio_01	**IDNUM**	4.85	5.401	62S021442H
Medical_audio_01	**STREET**	5.52	5.865	Yale
Medical_audio_01	**CITY**	6.451	6.732	Andergrove
Medical_audio_01	**STATE**	6.735	6.957	Tasmania
Medical_audio_01	**ZIP**	11.167	11.984	2042
Medical_audio_01	**MEDICALRECORD**	12.491	16.644	6270214.MFH
Medical_audio_01	**IDNUM**	14.464	15.782	62S02144

Table 3. Lists of teams whose Task 1 performance exceeded the baseline score of 0.138.

Table 3. Task 1 performance: teams exceeding baseline model (MER < 0.138)

TEAM_ID	Task1 Score	Task1 Rank	Final Rank
TEAM_7496	0.1293	3	1
TEAM_7882	0.1147	1	2
TEAM_7913	0.1299	4	3
TEAM_7870	0.1351	7	4
TEAM_7354	0.1330	5	5
TEAM_7757	0.1185	2	6
TEAM_7410	0.1337	6	7
TEAM_7372	0.1371	11	9
TEAM_7786	0.1357	8	10
TEAM_7373	0.1371	12	11
TEAM_7319	0.1368	10	17
TEAM_7901	0.1364	9	22

Table 4. Presents all teams whose Task 2 performance exceeded the baseline score of 0.3804.

TEAM_ID	Task2 Score	Task2 Rank	Final Rank
TEAM_7496	0.7103	1	1
TEAM_7882	0.5801	4	2
TEAM_7913	0.6051	2	3
TEAM_7870	0.6021	3	4
TEAM_7354	0.5060	7	5
TEAM_7757	0.4860	12	6

medRxiv preprint doi: https://doi.org/10.64898/2026.04.13.26349911; this version posted April 14, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

TEAM_7410	0.4870	11	7
TEAM_7251	0.5147	6	8
TEAM_7372	0.4952	10	9
TEAM_7786	0.4238	15	10
TEAM_7373	0.4784	13	11
TEAM_7342	0.5431	5	12
TEAM_7328	0.5015	8	13
TEAM_7897	0.4979	9	14
TEAM_7729	0.3901	17	15
TEAM_7764	0.4464	14	20

Table 5. Comparison of Whisper Model Performance Across Three Language Identification Strategies (language detect, language Token and English only Whisper.

Model Size	**Language Detect (MER)** (Overall MER / Eng MER / Ch MER)	**Specify Language Token (MER)** (Overall MER / Eng MER / Ch MER)	**English-only Whisper** (Overall MER / Eng MER / Ch MER)
Tiny	0.4422 / 0.3359 / 1.3992	0.4191 / 0.3359 / 1.1686	0.4924 / 0.2881 / 2.3311
Base	0.3031 / 0.2465 / 0.8123	0.3211 / 0.2465 / 0.9928	0.4978 / 0.2199 / 2.9992
Small	0.2348 / 0.1725 / 0.7955	0.2341 / 0.1717 / 0.7955	0.3399 / 0.1578 / 1.9793
Medium	0.1912 / 0.1362 / 0.6861	0.1902 / 0.1351 / 0.6861	0.2757 / 0.1405 / 1.4925
Large	0.1453 / 0.1326 / 0.2595	0.1426 / 0.1296 / 0.2595	
Large-v2	0.1474 / 0.1303 / 0.3015	0.1474 / 0.1303 / 0.3015	
Large-v3	0.1460 / 0.1238 / 0.3460	0.1456 / 0.1233 / 0.3460	
Large-v3-turbo	0.1488 / 0.1295 / 0.3220	0.1488 / 0.1295 / 0.3220	

Table 6. Comparison of Bilingual Joint and Independent Training Across Model Sizes

Model size	**Bilingual Joint Training**			**Bilingual Independent Training**		
	Overall MER	English MER	Chinese MER	Overall MER	English MER	Chinese MER
Tiny	0.3997	0.2911	1.3767	0.5266	0.3019	2.5495
Base	0.6325	0.6022	0.9047	0.2429	0.1857	0.7578
Small	0.1934	0.1688	0.4152	0.2060	0.1449	0.7556
Medium	0.3443	0.3025	0.7202	0.1394	0.1199	0.3155
Large	0.1752	0.1337	0.5481	0.1357	0.1121	0.3478
Large-v2	0.3357	0.2642	0.9798	0.1231	0.1077	0.2621
Large-v3	0.1567	0.1407	0.3009	0.1227	0.1101	0.2364
Large-v3-turbo	0.1508	0.1191	0.4363	0.1513	0.1309	0.3345

Table 7. The comparative performance of the models

Model	**Precision**	**Recall**	**F1 score**
Qwen 2.5 - 7B	0.8871	0.8372	0.8614
Llama 3.1 - 8B	0.8871	0.8470	0.8666
Mistral - 7B	0.8852	0.8391	0.8615
Med42 - 8B	0.8893	0.8329	0.8602
Gemma 2 - 9B	0.8572	0.7462	0.7979
OpenBioLLM-8B	0.8863	0.8299	0.8572